

Introduction to Data Science



Dr. José Ramón Iglesias

DSP-ASIC BUILDER GROUP
Director Semillero TRIAC
Ingeniería Electrónica
Universidad Popular del Cesar

Minería de Datos y el proceso de KDD



Vamos a trabajar en la preparación de los datos para obtener la ***“vista minable”***

Fase de Preparación de los Datos

- La información almacenada siempre tiene
 - ▮ Datos faltantes
 - ▮ Valores extremos
 - ▮ Inconsistencias
 - ▮ Ruido
- Tareas a realizar
 - ▮ Limpieza (ej: resolver outliers e inconsistencias)
 - ▮ Transformación (ej: discretización)

Limpieza de los datos

- ✦ En primer lugar, debe tenerse en cuenta que hay distintos tipos de variables o atributos.
- ✦ Para cada tipo se deberá realizar un análisis de sus valores.
 - Luego, se procederá a limpiarlos
 - ▮ Eliminando los valores con ruido
 - ▮ Determinando qué hacer con los valores faltantes.
 - ▮ Eliminando inconsistencias

Limpieza - Variables con ruido

- Las variables con ruido tendrán valores que caen fuera del rango de sus valores esperados llamados **outliers**.
- Por qué se originan?
 - ▮ Error humano en la carga de datos (ej: una persona puede aparecer con una altura de 5 metros).
 - ▮ Determinados cambios operacionales no han sido registrados en el proceso.

Es preciso analizar los **metadatos**

Limpieza - Valores faltantes

- Qué hacer con los valores nulos?
 - ▮ Ignorar la tupla.
 - ▮ Rellenar la tupla manualmente.
 - ▮ Usar una constante global para rellenar el valor nulo.
 - ▮ Utilizar el valor de la media u otra medida de centralidad para rellenar el valor.
 - ▮ Utilizar el valor de la media u otra medida de centralidad de los objetos que pertenecen la misma clase.
 - ▮ Utilizar alguna herramienta de Minería de Datos para calcular el valor más probable.



Hay datos faltantes
¿cómo los completamos?

Turbo Prep

More ▾

Find data, operators...etc



All Studio ▾

Result History

Exam



Data



Statistics



Visualizations



Annotations

Name



Type

Missing ▾

St...

Filter (12 / 12 attributes):

Search for Attributes



synopsis

Polynomial

0

Youthful [...] ary » (1)

A crimin [...] ents. (2)

A



release

Polynomial

4

Least

May (8)

Most

December (27)

Val

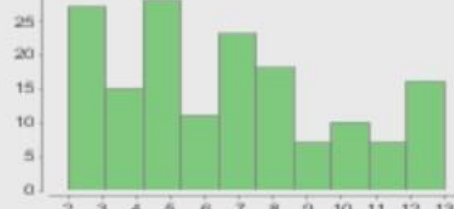
D



nominati...

Integer

16



Open visualizations

Min

2

Max

13



genre2

Polynomial

35

Least

War (1)

Most

Drama (74)

Val

D



Showing attributes 1 - 12



Examples: 178 Special Attributes: 0 Regular Attributes: 12

Repository



+ Import Data



▶ Training Resources (conn

▶ Samples

▶ Community Samples (cor

▶ DB

▶ ATAQUE_REDES (Laura)

▶ Local Repository (Laura)

▶ MBBS 2018 (Laura)

▶ MD (postgrado) (Laura)

▶ MD - PRACTICAS (Laura)

▶ MD_Educacion (Laura)

▶ MIDUSI (profesor)

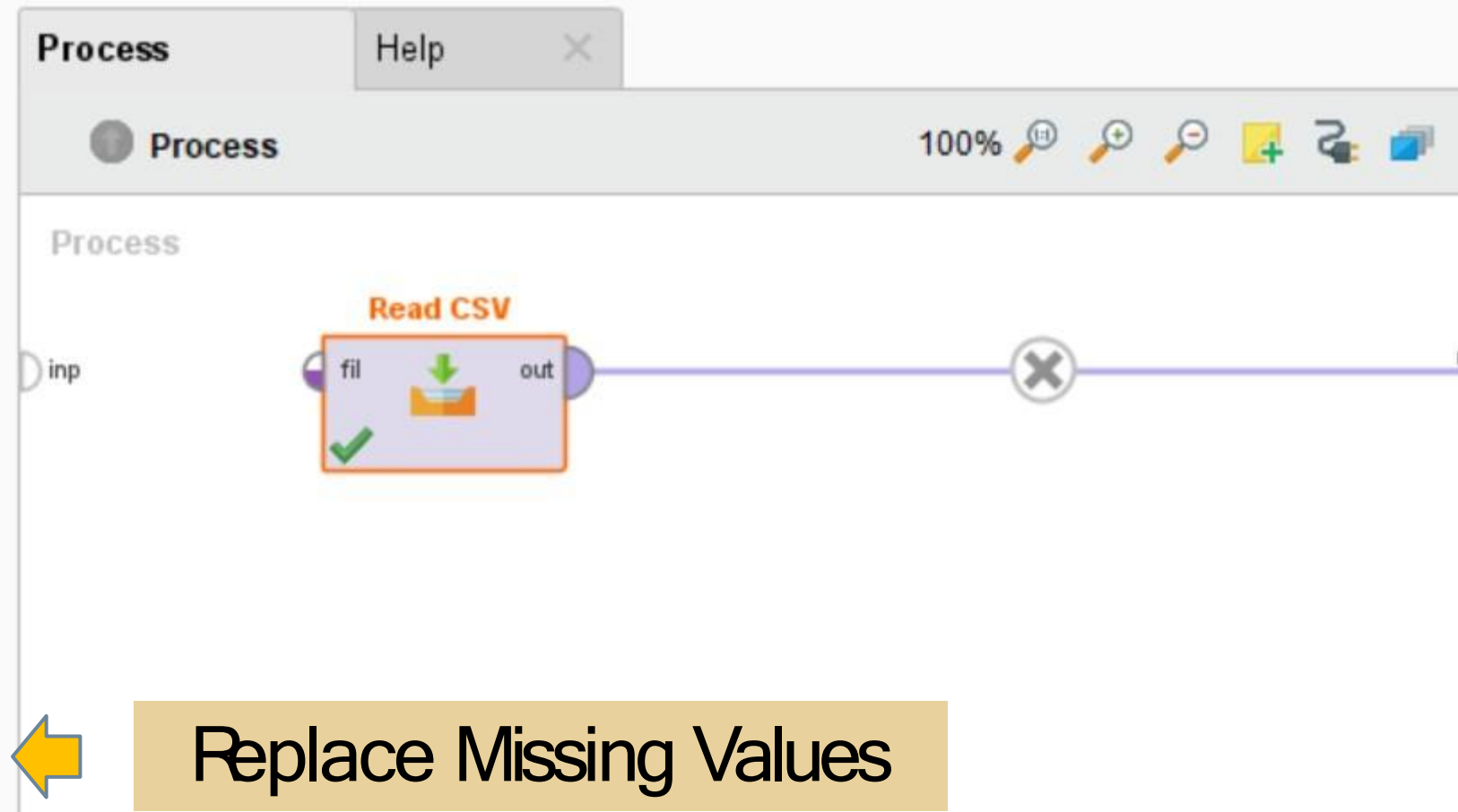
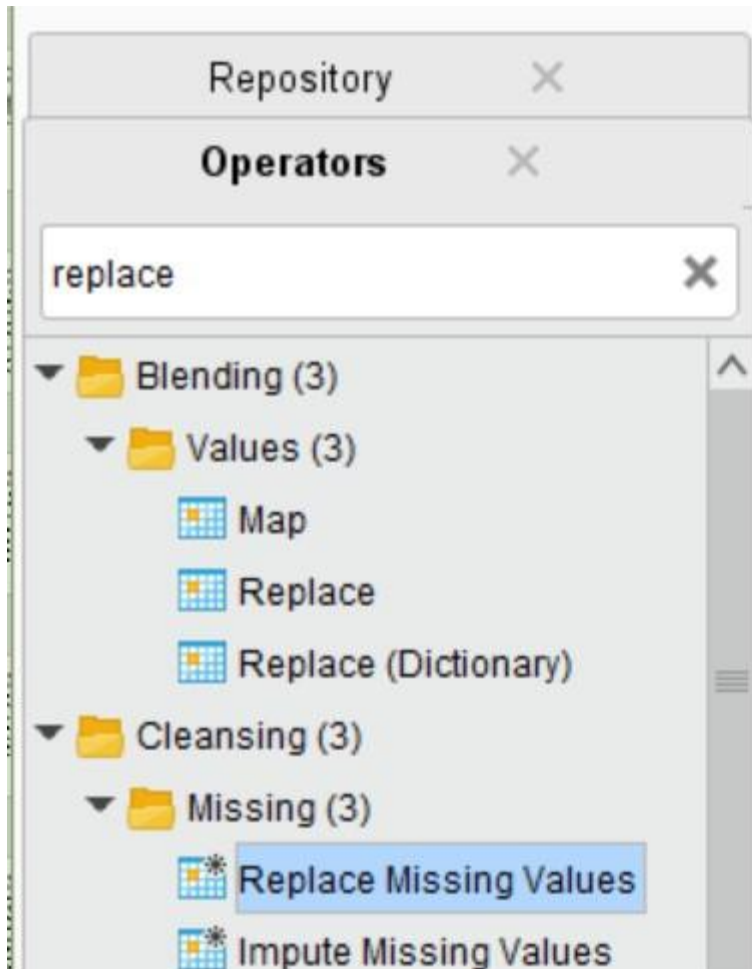
▶ MINERIA (Laura)

▶ Mineria2018 (Laura)

▶ Representacion2018 (Lau

▶ Tesina_AR (Laura)

Reemplazando los valores faltantes



Process

Process

Process

inp

Todos los atributos se completarán
con el promedio o la moda

Read CSV



Replace Missing Values



Edit Parameter List: columns



Edit Parameter List: **columns**
List of replacement functions for each column.

attribute	replace with
nominations	minimum
genre2	none



Add Entry



Remove Entry



Apply

Parameters

Replace Missing Values

☐ create view

attribute filter type

all

☐ invert selection

☐ include special attributes

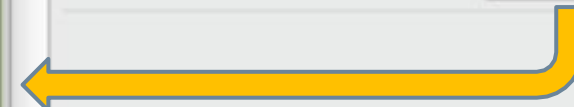
default

average

columns



Edit List (2)...



Se pueden indicar
reemplazos específicos

Recommen

Retriev

EJECUTE y verifique cómo
se completaron los atributos
nominations y release

File Edit Process View

Result History

Data

Statistics

Visualizations

Annotations

Name	Type	Missing	Filter (12 / 12 attributes):	
duration	Integer	0	69	238
genre1	Polynomial	0	Least Thriller (1)	Most Drama (86)
release	Polynomial	0	Least May (8)	Most December (31)
synopsis	Polynomial	0	Least Youthful [...] ary » (1)	Most A crimin [...] ents. (2)
genre2	Polynomial	35	Least War (1)	Most Drama (74)

Showing attributes 1 - 12

Examples: 178 Special Attributes: 0 Regular Attributes: 12

Repository

Import Data

- Training Resources (connected)
- Samples
- Community Samples (connected)
- DB
- ATAQUE_REDES (Laura)
- Local Repository (Laura)
- MBBS 2018 (Laura)
- MD (postgrado) (Laura)
- MD - PRACTICAS (Laura)
- MD_Educacion (Laura)
- MIDUSI (profesor)
- MINERIA (Laura)
- Mineria2018 (Laura)
- Representacion2018 (Laura)
- Tesina_AR (Laura)
- Tesis_AR (Laura)

Atributo GENRE1



Observe que hay muchos valores de la variable GENRE1 con muy pocos ejemplos


Atributo GENRE1

Vamos a reunir estas opciones como "OTROS"




Usaremos el operador MAP

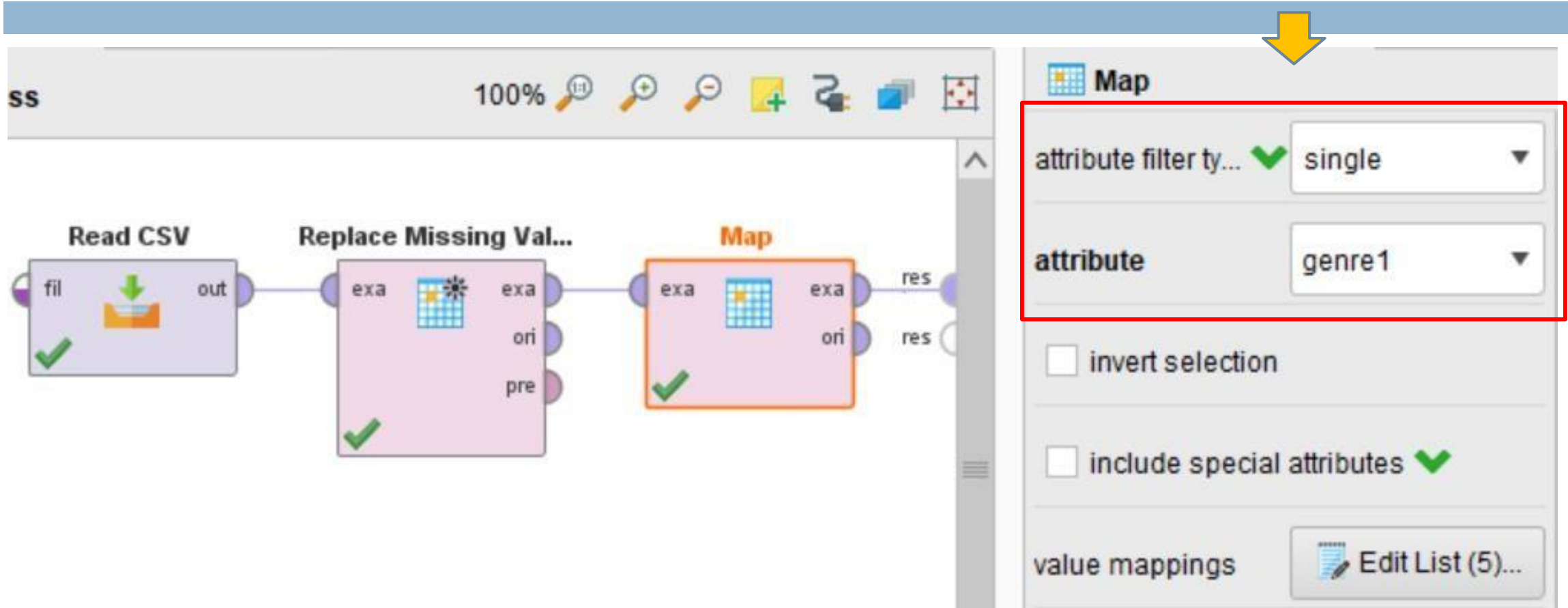
Nominal values



Index	Nominal value	Absolute count	Fraction
1	Drama	86	0.483
2	Biography	39	0.219
3	Comedy	24	0.135
4	Crime	16	0.090
5	Adventure	6	0.034
6	Action	3	0.017
7	Romance	2	0.011
8	Mystery	1	0.006
9	Thriller	1	0.006

 Close

Atributo GENRE1



The screenshot displays a data processing workflow in the Orange3 interface. The workflow consists of three widgets connected in sequence: 'Read CSV', 'Replace Missing Values', and 'Map'. The 'Map' widget is highlighted with an orange border and a green checkmark, indicating it is the active widget. A yellow arrow points from the top right towards the 'Map' widget's configuration panel.

The 'Map' widget configuration panel is shown on the right, with a red box highlighting the 'attribute filter type' and 'attribute' settings. The 'attribute filter type' is set to 'single' (indicated by a green checkmark), and the 'attribute' is set to 'genre1'.

Below the highlighted settings, there are two checkboxes: 'invert selection' (unchecked) and 'include special attributes' (checked, indicated by a green checkmark). At the bottom, there is a section for 'value mappings' with a button labeled 'Edit List (5)...'.

Atributo GENRE1

ss 100%

Read CSV

fil out

Replace Missing Val...

exa exa ori pre

Map

exa exa ori res res

Map

attribute filter ty... ✓ single

attribute genre1


☐ invert selection

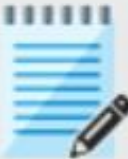
☐ include special attributes ✓

value mappings **Edit List (5)...**








Atributo GENRE1


 Edit Parameter List: value mappings

 Edit Parameter List: **value mappings**
The value mappings.

old values	new value
Action	OTRO
Romance	OTRO
Mystery	OTRO
Thriller	OTRO
Adventure	OTRO


 **Add Entry**  **Remove Entry**  **Apply**  **Cancel**


 **Map**

attribute filter ty...  **single** ▼

attribute **genre1** ▼

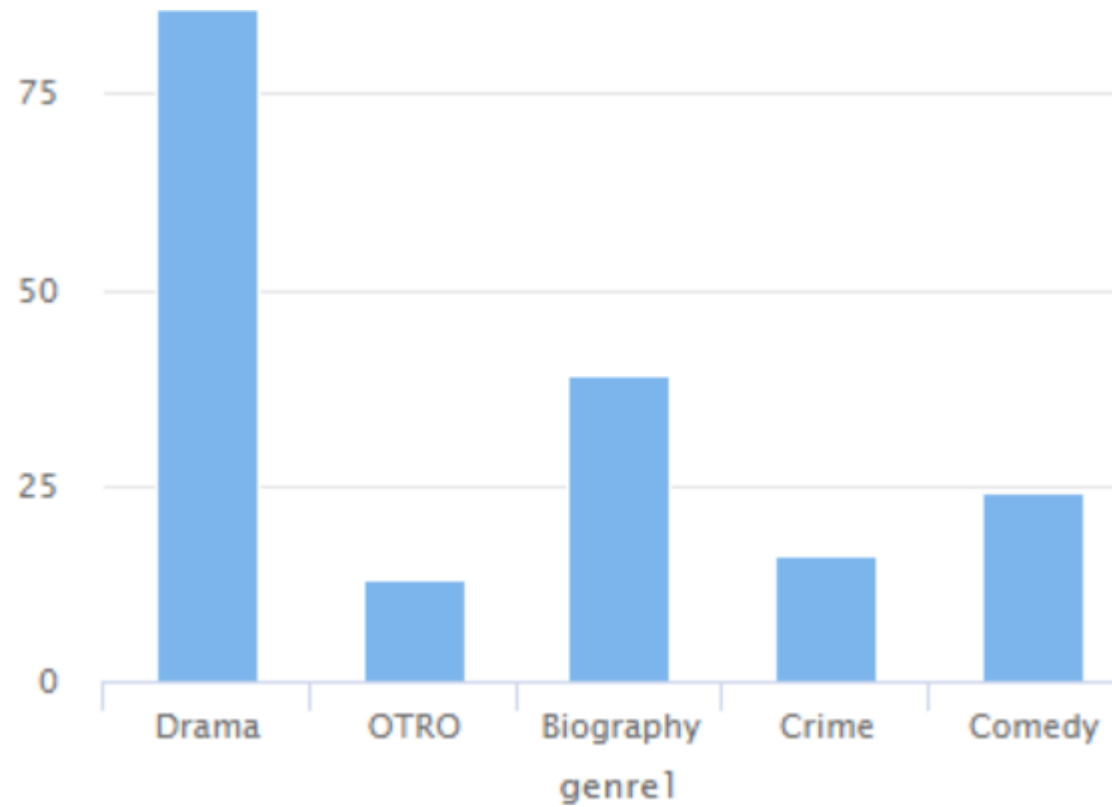
☐ invert selection

☐ include special attributes 


value mappings  **Edit List (5)...**



Atributo GENRE1



Nominal values



...	Nominal v...	Absolute c...	Fraction
1	Drama	86	0.483
2	Biography	39	0.219
3	Comedy	24	0.135
4	Crime	16	0.090
5	OTRO	13	0.073

Close

Atributo GENRE1

En lugar de mapear cada valor a OTRO se puede utilizar una expresión regular

The screenshot displays an Alteryx workflow with three tools: 'Read CSV', 'Replace Missing Val...', and 'Map'. The 'Map' tool is highlighted, and its configuration panel is open on the right. The configuration shows the 'attribute' set to 'genre1' and the 'replace what' field containing the regular expression '|Romance|Mystery|Thriller'. The 'replace by' field is set to 'Otro', and the 'consider regular expressions' checkbox is checked. A blue arrow points from the 'replace what' field to the regular expression text. Another blue arrow points from the 'consider regular expressions' checkbox to the same text. Below the workflow, a light blue box contains the text 'Adventure|Action|Romance|Mystery|Thriller'.

Read CSV

Replace Missing Val...

Map

100%

attribute filter type ☒ single

attribute genre1

☐ invert selection

☐ include special attributes ☒

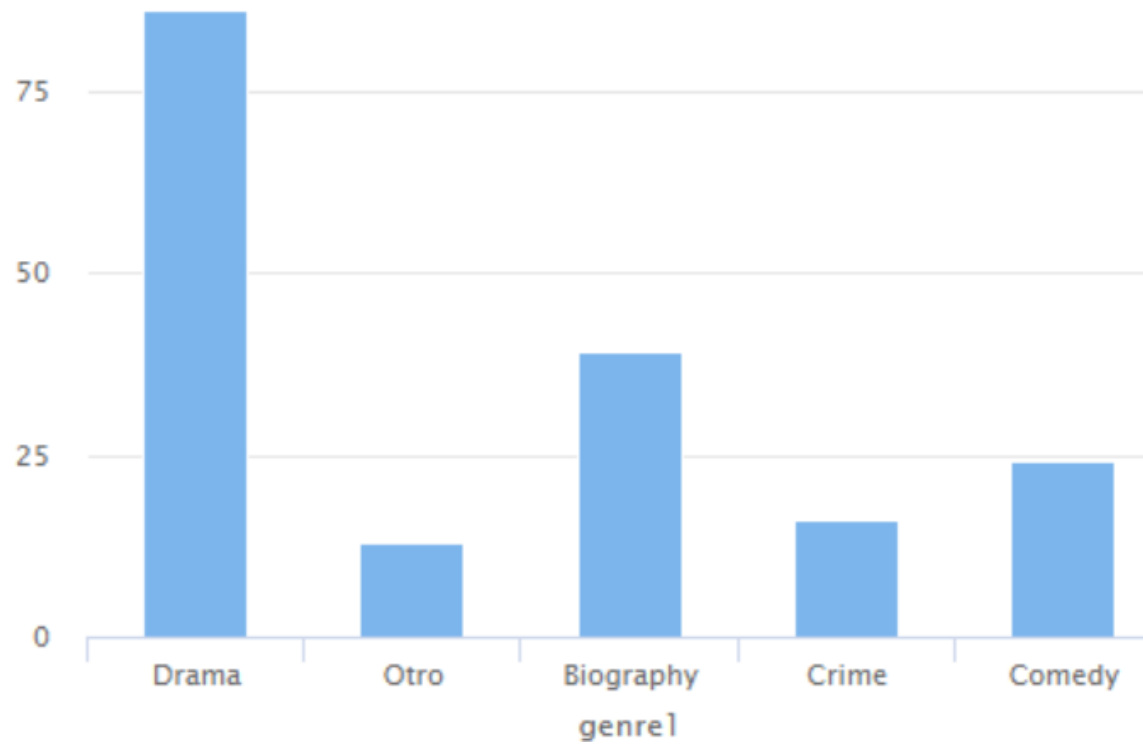
replace what ☒ |Romance|Mystery|Thriller

replace by Otro


☒ consider regular expressions

Adventure|Action|Romance|Mystery|Thriller

Atributo GENRE1



Nominal values



Ind...	Nominal value	Absolute count	Fraction
1	Drama	86	0.483
2	Biography	39	0.219
3	Comedy	24	0.135
4	Crime	16	0.090
5	Otro	13	0.073

Close

Transformación de atributos

- Es una de las etapas más importantes porque de ella depende el éxito del proceso.
- Los atributos serán transformados según las necesidades del algoritmo a aplicar.
- Es probable que deban derivarse variables nuevas.
- También es posible que se reduzcan variables convirtiéndolas en información más significativa.

Transformación de atributos

- **Según el algoritmo a aplicar**, las transformaciones más habituales son:
 - ▮ Reducción de dimensionalidad
 - ▮ Aumento de dimensionalidad
 - ▮ Discretización de atributos numéricos
 - ▮ Numerización de atributos nominales
 - ▮ Normalización de atributos

Transformación de atributos

□ Reducción de dimensionalidad

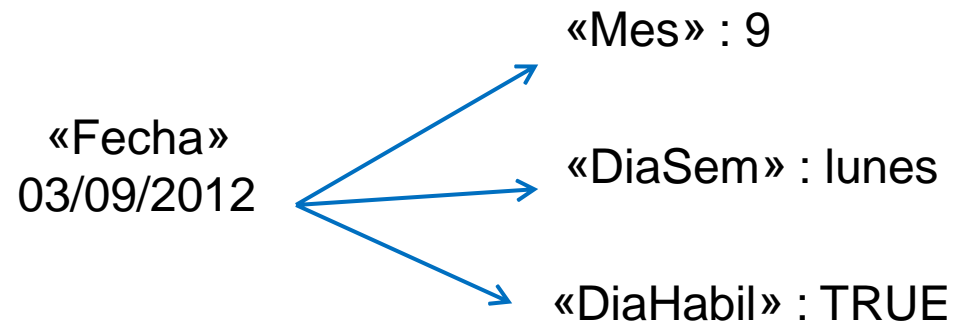
- ▮ Cambia el espacio de entrada por otro que tiene menor dimensión.
- ▮ Se busca mejorar la relación entre la cantidad de ejemplos y la cantidad de atributos.

▮ Ejemplos

- Análisis de componentes principales (PCA)
- Red SOM (self-organizing maps)

Transformación de atributos

- Aumento de la dimensionalidad a través de la **creación de características**
 - Atributos numéricos : se utiliza suma, resta, producto, división, máximo, mínimo, media, cuadrado, raíz cuadrada, seno, coseno, etc.
 - Fechas: brindan poca información si se las usa directamente.



Transformación de atributos

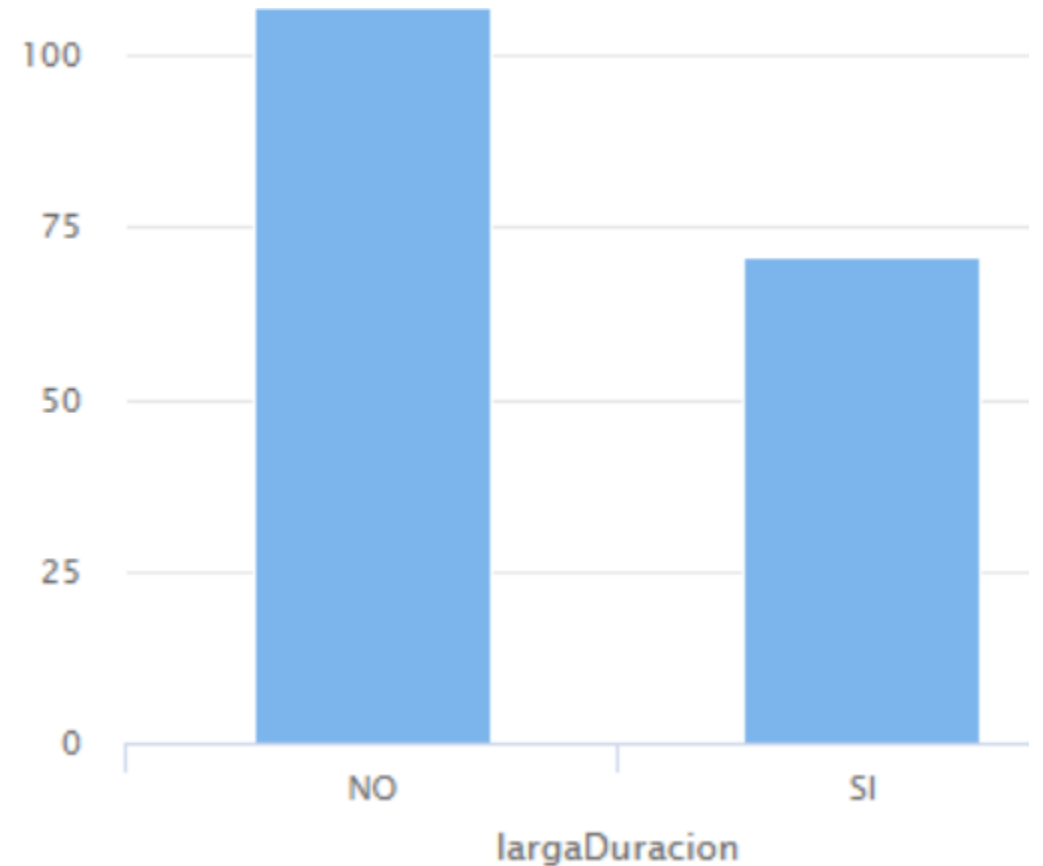
- Aumento de la dimensionalidad a través de la **creación de características**
 - Atributos nominales:
 - Se utilizan las operaciones lógicas, igualdad o desigualdad, condiciones **M-de-N** (TRUE si al menos M de las N condiciones son verdaderas).
 - Se puede generar un valor numérico a partir de valores nominales, por ejemplo, las variables **X-de-N** (retorna el entero X de las N condiciones que son ciertas)

Ejemplo de creación de atributos

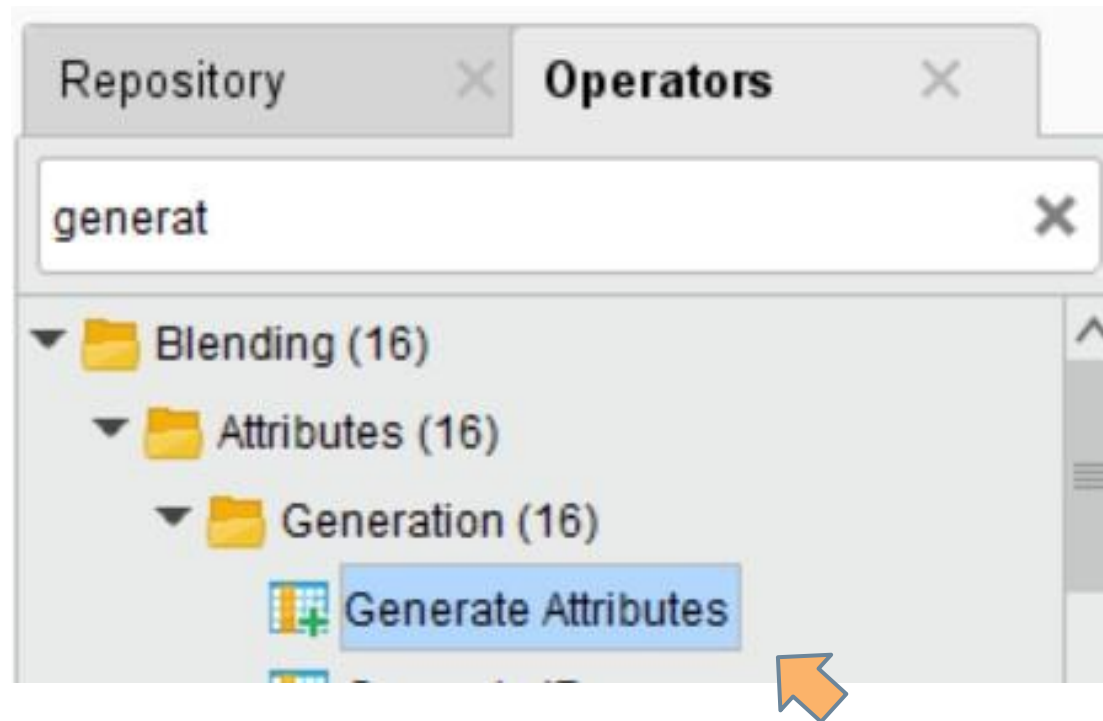
Atributo derivado	Fórmula
Índice de obesidad	$\text{Altura}^2 / \text{peso}$
Hombre familiar	Casado, varón e (hijos > 0)
Síntomas SARS	3-de-5 (fiebre alta, vómitos, tos, diarrea, dolor de cabeza)
Riesgo de póliza	X-de-N (edad<25, varón, años que conduce<2, vehículo deportivo)
Beneficios Brutos	Ingresos - Gastos
Beneficios netos	Ingresos - Gastos - Impuestos
Desplazamiento	Pasajeros * kilómetro
Duración media	Segundos de llamada / número de llamadas
Densidad	Población / Area
Retardo compra	Fecha compra - Fecha campaña

Ejercicio

- Genere un nuevo atributo **largaDuracion** cuyo valor será “SI” si la película tiene una duración superior a 2 horas y “NO” en caso contrario.
- Grafique este nuevo atributo utilizando un diagrama de barras.



Generando un nuevo atributo



Generemos un nuevo atributo
utilizando el componente
Generate Attributes

Generando un nuevo atributo

□ Operador **Generate Attributes**

The screenshot displays the Orange3 data mining software interface. On the left, a workflow is visible with four operators: 'Read CSV', 'Replace Missing Val...', 'Map', and 'Generate Attributes'. The 'Generate Attributes' operator is highlighted with an orange border. On the right, the configuration panel for the 'Generate Attributes' operator is shown. It includes a 'function description...' field with an 'Edit List (1)...' button, a 'keep all' checkbox, and an orange arrow pointing to the 'Edit List (1)...' button. A yellow callout box with a red border contains the text: 'Antes de ejecutarlo haga click aquí para configurarlo'.

100%

Read CSV

fil out

Replace Missing Val...

exa exa ori pre

Map

exa exa ori

Generate Attributes

exa exa ori res res

Generate Attributes

function description... Edit List (1)...

☒ keep all

Antes de ejecutarlo haga click aquí para configurarlo

Generación de un nuevo atributo

Edit Parameter List: function descriptions

Edit Parameter List: **function descriptions**
List of functions to generate.

attribute name	function expressions
<input type="text" value="largaDuracion"/>	<input si\",\"no\")"="" type="text" value="if(duration>120,\"/>

Nombre del nuevo atributo

definición


Remove E... Cancel

Generación de un nuevo atributo

Edit Parameter List: function descriptions

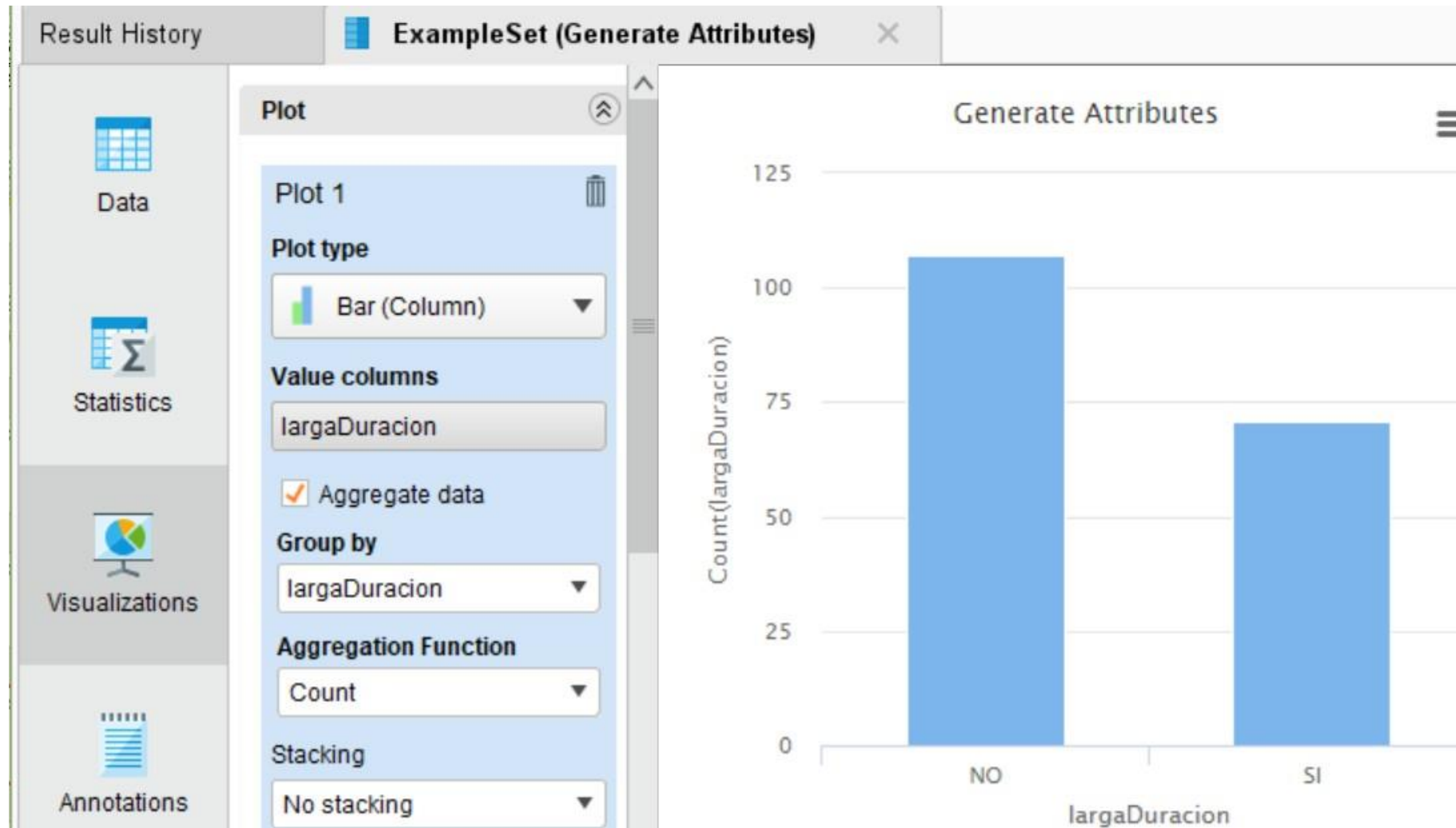
Edit Parameter List: **function descriptions**
List of functions to generate.

attribute name	function expressions
largaDuracion	if(duration>120,"SI","NO")

 Add

if (duration>120 , "SI" , "NO")

Diagrama de barras del atributo generado



Transformación de atributos

□ DISCRETIZACION

- ▮ Algunos algoritmos de minería de datos sólo operan con atributos cualitativos. La discretización convierte los atributos numéricos en ordinales.

□ NUMERIZACION

- ▮ Es el proceso contrario a la discretización. Convierte atributos cualitativos en numéricos.

□ NORMALIZACION

- ▮ Permite expresar los valores de los atributos sin utilizar las unidades de medida originales facilitando su comparación y uso conjunto.

Discretización

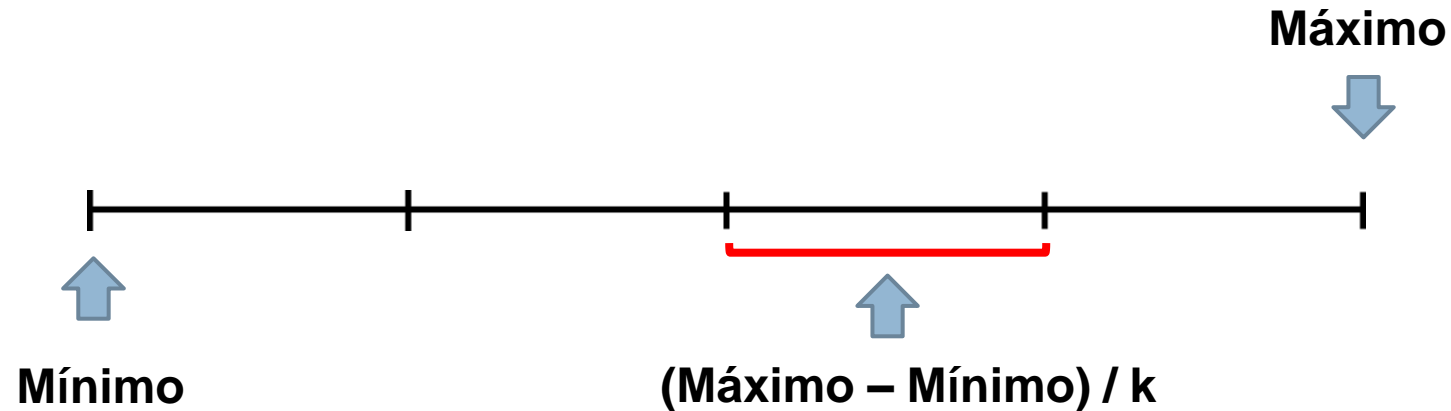
- Convierte un valor numérico en un nominal ordenado (que representa un intervalo o "*bin*")
- **Ejemplo:** Podemos transformar
 - ▮ la edad de la persona en categorías: $[0, 12]$ niño, $(12-21)$ joven, $[21, 65]$ adulto y >65 anciano.
 - ▮ La calificación de un alumno en: $[4, 10]$ aprobado o $[0, 4)$ desaprobado

Discretización

- Puede discretizarse en un número fijo de intervalos. El ancho del intervalo se calcula
 - ▮ Dividiendo el rango en partes iguales
 - ▮ Dividiendo la cantidad de ejemplos en partes iguales (igual frecuencia)
 - ▮ Indicando los límites de cada intervalo en forma manual.
- Averigüe por otras variantes de discretización

Discretización por rango

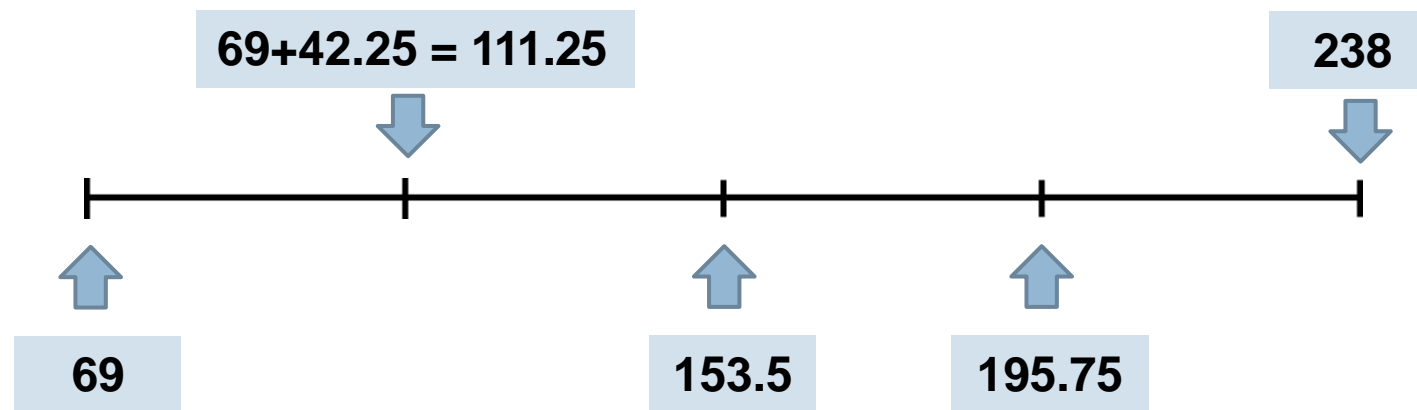
- El objetivo es dividir el rango del atributo (intervalo entre el máximo y el mínimo) en una cierta cantidad k de partes iguales.
- Los valores comprendidos en una misma parte serán asociados al mismo valor ordinal.
- Ejemplo: $k=4$



Discretización por rango

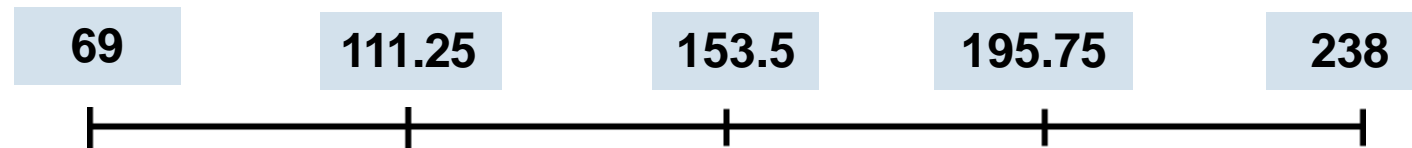
□ Ejemplo: Discretizar el atributo DURATION en 4 intervalos de igual longitud

- ▮ DURATION toma valores entre 69 y 238 minutos. Si dividimos el rango en 4 partes iguales, cada una tendría una longitud de $(238-69)/4 = 42.25$



Discretización por rango

- **Ejemplo: Discretizar el atributo DURATION en 4 intervalos de igual longitud**



Valor	Intervalo	Frecuencia
range1	$[-\infty - 111.25]$	75
range2	$(111.25 - 153.5]$	86
range3	$(153.5 - 195.75]$	15
range4	$(195.75 - \infty]$	2

Discretización por rango

- DURATION discretizado en 4 intervalos de igual longitud

The screenshot shows a data processing workflow on the left and the configuration panel for the 'Discretize' operator on the right.

Workflow:

- A 'Premios.csv' file is loaded into a 'fil' (file) port of a 'Discretize' operator.
- The 'Discretize' operator has an 'out' (output) port connected to a 'res' (result) port.
- The 'Discretize' operator is highlighted with an orange border.
- A blue arrow points from the text box below to the 'Discretize' operator.

Discretize (Discretize by Binning) Configuration:

- ☐ create view
- attribute filter type: single
- attribute: duration
- ☐ invert selection
- ☐ include special attributes
- number of bins: 4

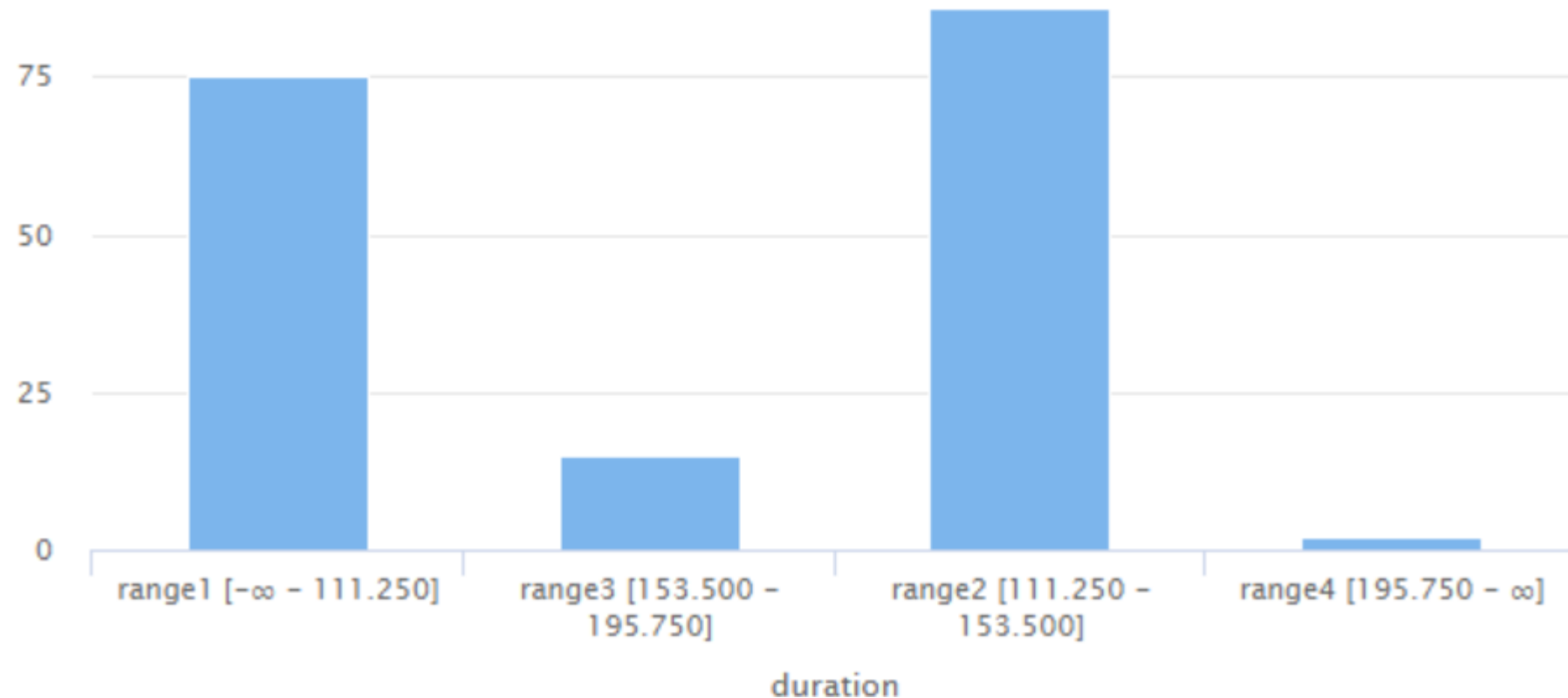
A red arrow points from the text box on the right to the 'number of bins' field.

Vamos a discretizar el atributo **DURACION** utilizando el operador **Discretize by Binning**

Indicar la cantidad de intervalos

Discretización por rango

- DURATION discretizado en 4 intervalos de igual longitud



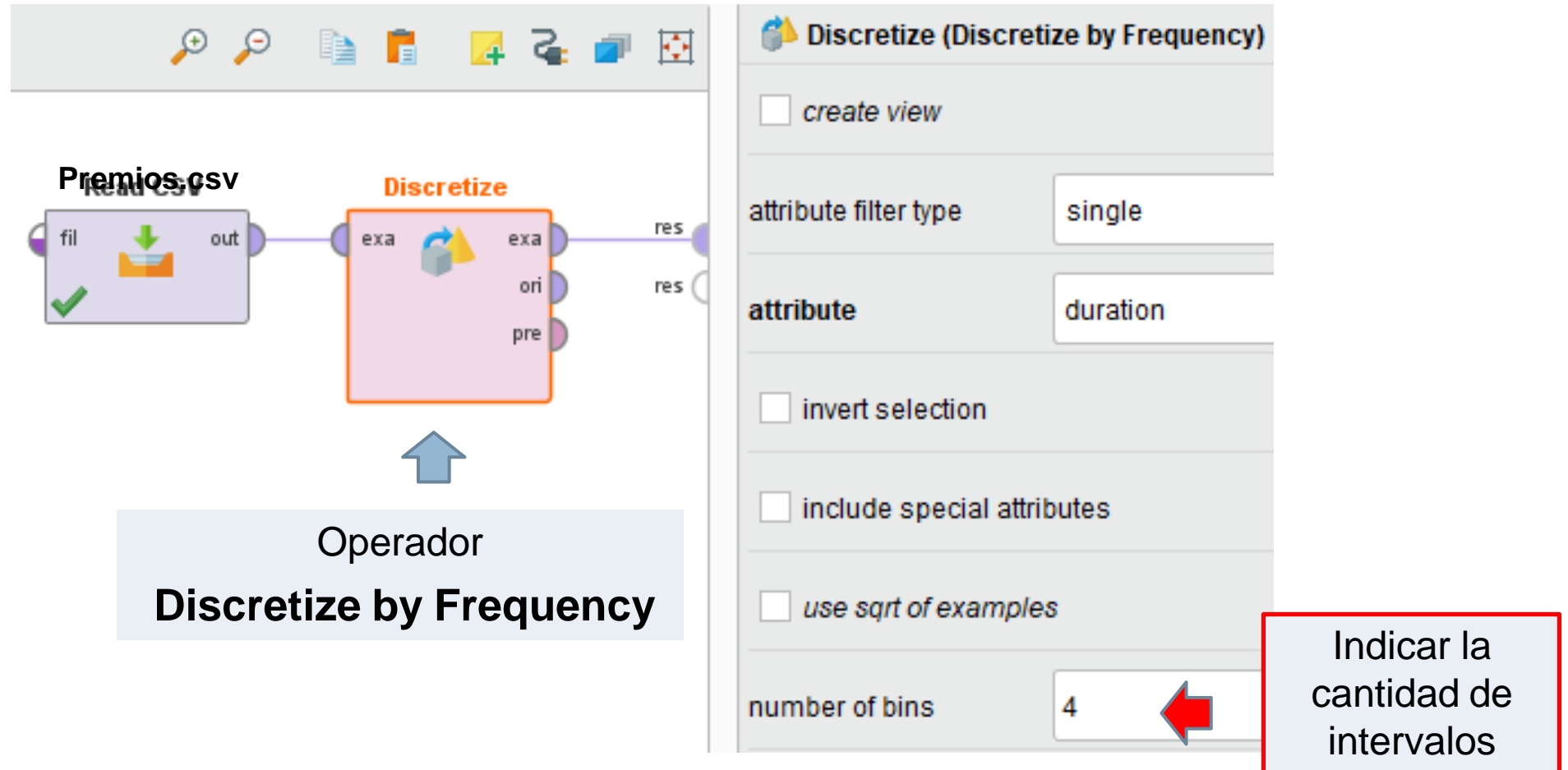
Discretización por frecuencia

- El objetivo es dividir los valores del atributo numérico en k partes con la misma cantidad de valores en cada una de ellas.
- Nótese que el atributo debe tener al menos k valores diferentes.
- **Ejemplo: Discretizar DURATION en 4 intervalos de igual frecuencia**

Valor	Intervalo	Frecuencia
range1	$[-\infty - 104.5]$	45
range3	$(115.5 - 130]$	45
range2	$(104.5 - 115.5]$	44
range4	$(130 - \infty]$	44

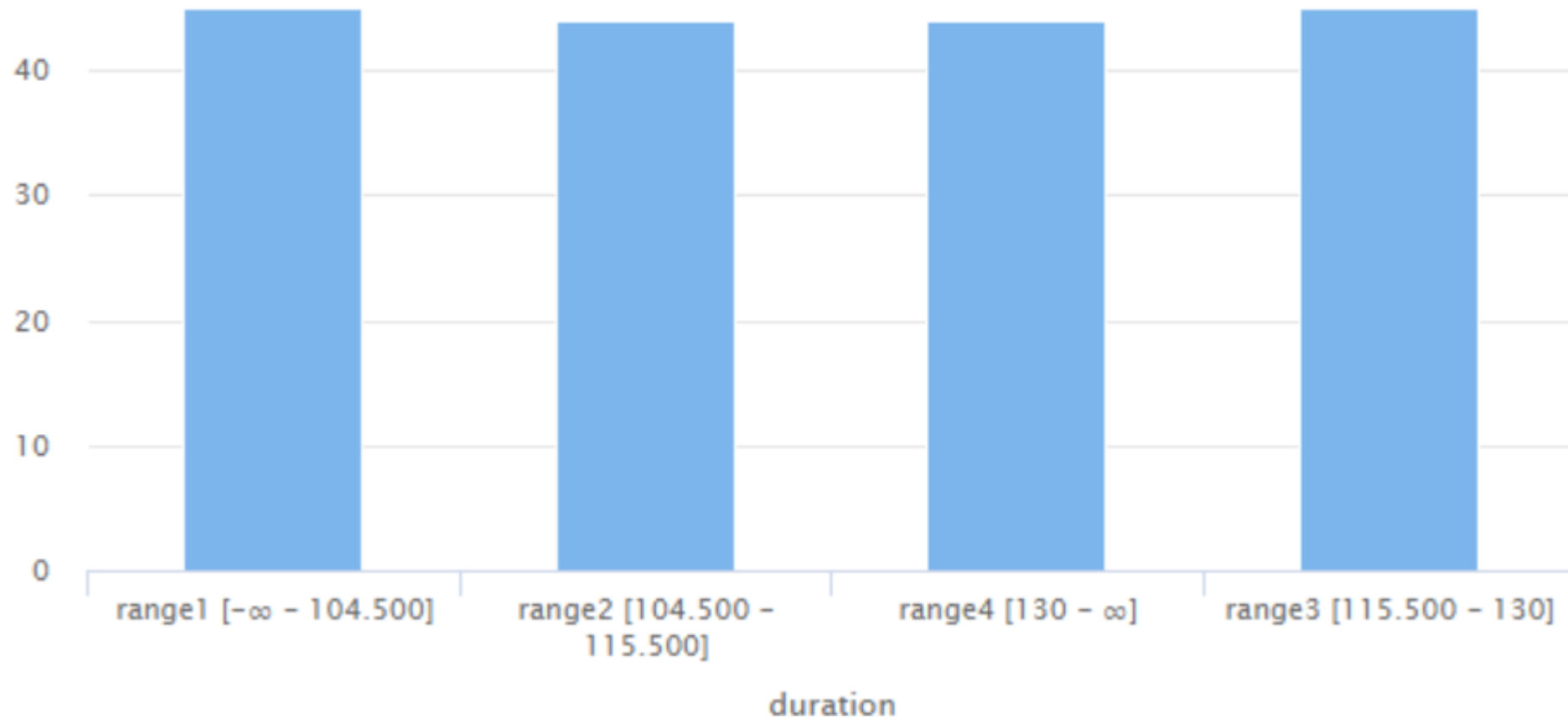
Discretización por frecuencia

- DURATION discretizado en 4 intervalos de igual frecuencia



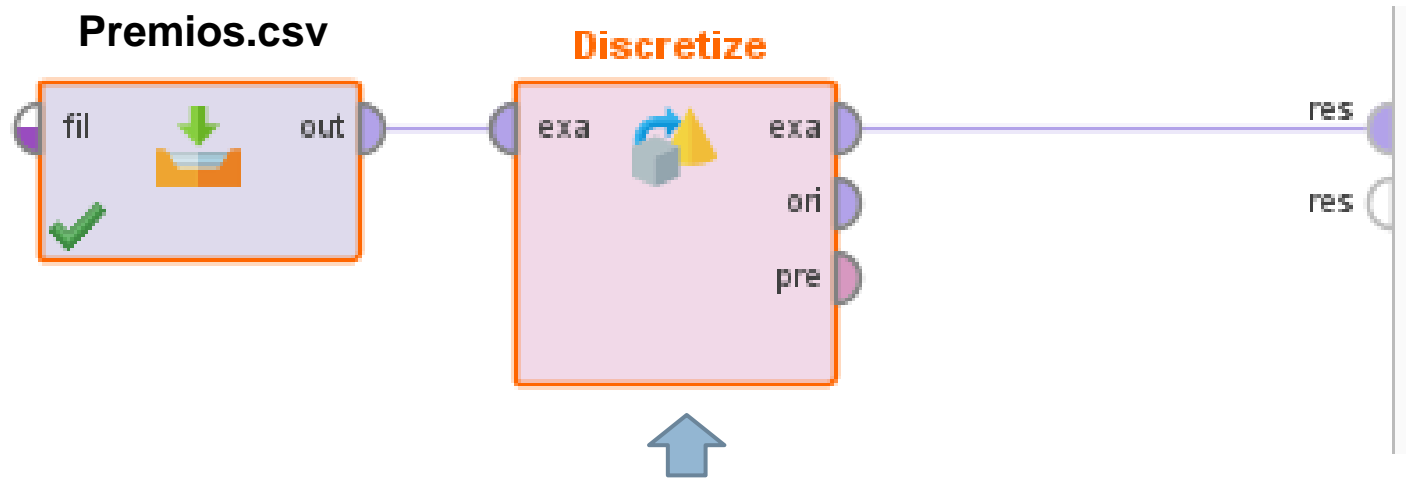
Discretización por frecuencia

- DURATION discretizado en 4 intervalos de igual frecuencia



Discretización especificada por el usuario

- Se indican los umbrales a utilizar en forma manual

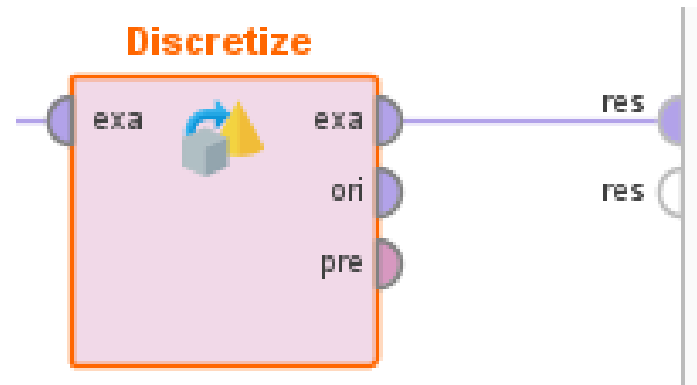


Vamos a discretizar el atributo ***DURACION***
utilizando el operador

Discretize by User Specification

Discretización especificada por el usuario

□ Operador **Discretize by User Specification**



Se selecciona el atributo
DURATION

Parameters

Discretize (Discretize by User Specification)

☐ create view

attribute filter type single

attribute duration

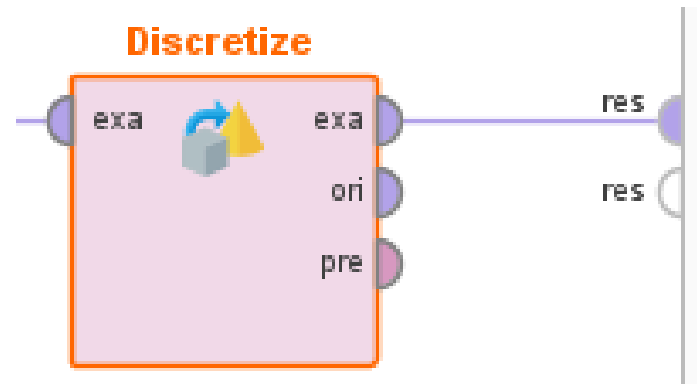
☐ invert selection

☐ include special attributes

classes [Edit List \(2\)...](#)

Discretización especificada por el usuario

□ Operador **Discretize by User Specification**



Aquí se indican los intervalos


The screenshot shows the "Parameters" dialog box for the "Discretize (Discretize by User Specification)" operator. The dialog has a title bar with a close button. The main content area contains the following settings:


- ☐ *create view*
- attribute filter type: **single**
- attribute: **duration**
- ☐ *invert selection*
- ☐ *include special attributes*
- classes: **Edit List (2)...**

The "classes" section is highlighted with a red border.




Discretización especificada por el usuario

□ Operador **Discretize by User Specification**


 Edit Parameter List: classes

 Edit Parameter List: **classes**
Defines the classes and the upper limits of each class.

class names	upper limit
BREVE	100.0
NORMAL	136.0
LARGA	Infinity

 Add Entry  Remove Entry  Apply

Parameters

 **Discretize (Discretize by User Specification)**


☐ create view

attribute filter type single

attribute duration

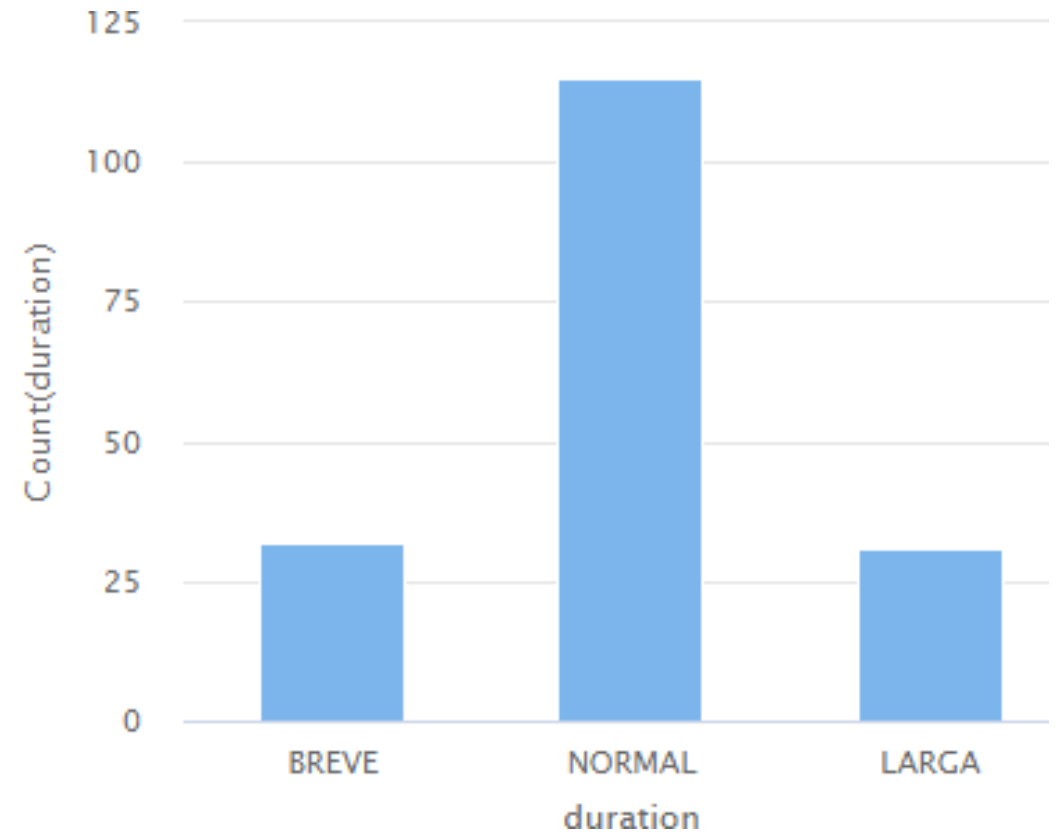
☐ invert selection

☐ include special attributes

classes  Edit List (2)...

Discretización especificada por el usuario

- Si $DURATION \leq 100$, BREVE
- Si $(DURATION > 100)$ y $(DURATION \leq 136)$, NORMAL
- Si $(DURATION > 136)$, LARGA



Numerización

- En ocasiones los atributos nominales u ordinales deben convertirse en números.
- Para los nominales suele utilizarse una representación binaria y para los ordinales suele utilizarse una representación entera.
- Es importante considerar que si se numeran en forma correlativa los valores de un atributo nominal se agrega un orden que originalmente no está presente en la información disponible.

Numerización Binaria (dummy)

- La numerización binaria reemplaza al atributo nominal por tantos atributos numéricos binarios como valores distintos pueda tomar.
- Las denominaciones de estos nuevos atributos surgen de igualar el nombre original con cada uno de los posibles valores.
- Para un mismo ejemplo sólo uno de estos nuevos atributos tendrá valor 1 y el resto 0.



Views:

Design

Results

Turbo Prep

Auto Model

Find data, operators...etc

All Studio

Repository

Operators

nomin

Blending (11)

Attributes (10)

Types (10)

- Numerical to Binominal
- Numerical to Polynominal
- Nominal to Binominal
- Nominal to Text
- Nominal to Numerical
- Nominal to Date
- Text to Nominal
- Date to Nominal
- Parse Numbers
- Guess Types

Process

Help

Process

100%

Process

Read CSV

Replace Missing Val...

Nominal to Numerical

Nominal to Numerical

Parameters

Nominal to Numerical

☐ create view

attribute filter ... ☒ single

attribute Sex

☐ invert selection

☐ include special attributes

coding type dummy codi...

☐ use comparison groups

unexpected valu... all 0 and wa...

☐ use underscore in name

[Hide advanced parameters](#)

[Change compatibility \(9.2.000\)](#)

Recommended Operators

Retrieve

68%

Select Attributes

48%

Set Role

42%

Numerización Binaria de SEX

Row No.	Sex = M	Sex = F	Year	Age	Actor	Film	nominatio
1	1	0	1928	44	Emil Jannings	The Last Co...	2
2	0	1	1928	22	Laura Gainor ...	Sunrise	5
3	1	0	1929	38	Warner Baxter	In Old Arizona	5
4	0	1	1929	37	Mary Pickford	Coquette	2
5	1	0	1930	62	George Arliss	Disraeli	3
6	0	1	1930	30	Norma Shear...	The Divorcee	4
7	1	0	1931	53	Lionel Barry...	A Free Soul	3
8	0	1	1931	62	Marie Dressler	Min and Bill	2
9	1	0	1932	41	W. Beery(47)/...	The Champ/...	4
10	0	1	1932	32	Helen Hayes	Sin of Madelon	2

Normalización

- Se aplica según el modelo que se va a construir.
- La más común es la **normalización lineal uniforme**

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- Es muy sensible a valores fuera de rango (outliers).
- Si se recortan los extremos se obtiene valor negativos y/o mayores a 1.

Normalización

- Existen otras transformaciones. Por ejemplo, si los datos tienen distribución normal se pueden **tipificar**

$$X' = \frac{X - \text{media}(X)}{\text{desviacion}(X)}$$

- De esta forma los datos se distribuyen normalmente alrededor de 0 con desviación 1.

Normalización - Operador Normalize

The screenshot displays a data workflow interface with three operators: 'Read CSV', 'Replace Missing Val...', and 'Normalize'. The 'Normalize' operator is highlighted with an orange border. The right-hand panel shows the configuration for the 'Normalize' operator.

100% [Zoom icons]

Read CSV
fil [icon] out

Replace Missing Val...
exa [icon] exa
ori
pre

Normalize
exa [icon] exa
ori
pre

res
res

Normalize

- ☐ create view
- attribute filter ty... ☒ all
- ☐ invert selection
- ☐ include special attributes
- method ☒ Z-transformat...

Z-transformat...
Z-transformation
range transformation
proportion transformation
interquartile range

Resumen

PREPARACION DE LOS DATOS

- Completar datos faltantes
- Operador MAP
- Generación de características o atributos nuevos
- Transformaciones
 - ▮ Discretización por rango, por frecuencia e indicada por el usuario
 - ▮ Numerización: codificación entera y codificación binaria
 - ▮ Normalización: Lineal y Estandarización